KNOWLEDGE CONSTRUCTION: THE ROLE OF DATA MINING TOOLS

Maribel Yasmina Santos Isabel Ramos

{maribel,iramos}@dsi.uminho.pt

Information Systems Department
University of Minho
Azurém Campus
4800-058 Guimarães
Portugal

Tel.: +351 253 510317/19
Fax: +351 253 510300

Keywords: data mining, knowledge discovery, knowledge construction, organisational knowledge.

Stream: Co-ordination - Knowledge Management

Abstract

This paper seeks to integrate the process of knowledge discovery in databases in the wider context of the creation and sharing of organisational knowledge. The focus on the process of knowledge discovery has been mainly technological. The paper attempts to enrich that perspective by stressing the insights gained by integrating the knowledge discovery process into the social process of knowledge construction that makes KDD meaningful. In order to achieve this goal, a test case is presented. A component of the database of the Portuguese Army was used to test the PADRÃO system. This system integrates a set of databases and principles of qualitative spatial reasoning, which are implemented in the Clementine Data Mining system. The process and the results obtained are then discussed in order to stress the insights that emerge when the focus changes from technology to the social construction of knowledge.

1 Introduction

The process of Knowledge Discovery in Databases automates the discovery of relationships and other descriptions from data. Data Mining refers to the application algorithms for extracting patterns from data without the additional steps of the knowledge discovery process, such as incorporating appropriate prior knowledge and interpretation of results.

The patterns and/or rules extracted in the knowledge discovery process are considered to be knowledge in terms of the theoretical concepts behind them. This paper presents another view in which it is argued that theses rules cannot be considered knowledge, since knowledge can reside only in human minds, which are continuously growing in relation to their inner and outer worlds. Instead of knowledge, these patterns and/or rules are considered to be knowledge representations or information.

The interpretation of the data mining results does not lead immediately to knowledge. The developed understanding needs to be tried out in organisational situations such as decision-making processes, organisational projects and the reformulation of practices. In this way, organisational actors create new ideas and mental models, reshape previous ones, redefine conceptual linkages and develop emotional responses towards ideas and models, i.e. organisational actors create new experiences and relationships with the surrounding reality.

In this paper Section 2 looks at an overview of the process of knowledge discovery and presents its several phases. Section 3 describes a test case in which a component of the database of the Portuguese Army is analysed using the principles of the knowledge discovery process. Section 4 then discusses some general issues related to the results obtained in the relevant test case and Section 5 concludes the paper with some comments and remarks about future work.

2 Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is a complex process concerning the discovery of relationships and other descriptions from data. Data Mining refers to the application algorithms used to extract patterns from data without the additional steps of the KDD process e.g. the incorporation of appropriate prior knowledge and the interpretation of results *(Fayyad and Uthurusamy 1996).*

Different tasks can be performed in the knowledge discovery process and several techniques can be applied for the execution of a specific task. Among the available tasks are *classification*, *clustering*, *association*, *prediction*, *estimation* and *summarisation*. KDD applications integrate a variety of Data Mining algorithms. The performance of each technique (algorithm) depends upon the task to be carried out, the quality of the available data and the objective of the discovery. The most popular Data Mining algorithms include *neural networks*, *decision trees*, *association rules* and *genetic algorithms (Han and Kamber 2001).*

The steps (Figure 1) of the KDD process include data selection, data treatment, data pre-processing, data mining and interpretation of the results. This process is

interactive, because it requires user participation, and iterative, because it allows for going back to a previous phase and then proceeding with the knowledge discovery process. The steps of the KDD process are briefly described below.

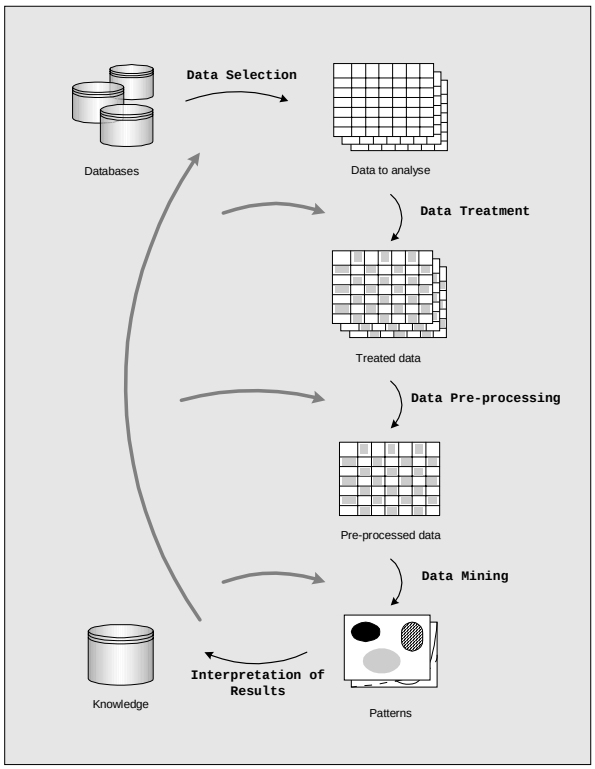

Figure 1 – The steps of the knowledge discovery process

Data Selection.
This step allows for the selection of relevant data needed for the execution of a defined Data Mining task. In this phase the minimum sub-set of data to be selected, the size of the sample needed and the period of time to be considered must be evaluated.

Data Treatment.
This phase is concerned with the cleaning up of selected data, allowing for the treatment of corrupted data and the definition of strategies for dealing with missing data fields.

Data Pre-Processing.
This step makes possible the reduction of the sample destined for analysis. Two tasks can be carried out here: i) the reduction of the number of rows or, ii) the reduction of the number of columns. In the reduction of the number of rows, data can be generalised according to the hierarchies or attributes of the defined domain with continuous values being transformed into discreet values depending on the defined classes. The reduction of the number of columns attempts to verify if any of the selected attributes can be omitted later.

Data Mining.
Several algorithms can be used for the execution of a given data mining task. In this step, several available algorithms are evaluated in order to identify the most appropriate for the defined task. The selected one is applied to the relevant data in order to find implicit relationships or other interesting patterns that exist in the data.

Interpretation of Results.
The interpretation of the discovered patterns aims at evaluating their utility and importance with respect to the application domain. In this step it may be determined that relevant attributes were ignored in the analysis, thus suggesting that the process should be repeated.

The following paragraphs are aimed at clarifying the notion of knowledge and knowledge discovery of the authors, which is considered to be knowledge creation supported by data mining tools.

Knowledge is seen as socially constructed through human action and interaction that happens in specific socio-cultural contexts. These contexts, organisational and societal, are responsible for stable patterns of action and interaction that contribute to the objective appearance of working realities. This appearance then contributes to the notion of an objective knowledge that can be stored outside the human mind or discovered independently of a frame of interpretation, as would be the case for knowledge discovery using data mining systems.

This view is challenged elsewhere by the second author of this paper *(Ramos and Carvalho)*. For the purpose of this paper, however, the following aspects of the process illustrated in Figure 1 are highlighted:

1. The facts, events, things, concepts, models, ideas registered and stored in databases cannot be considered knowledge, since knowledge can reside only in human minds continuously growing in connection with the inner and outer worlds. Stored and manipulated data are representations of knowledge or information.
2. All meaningful information, which fits human mental and social spaces, has potential for creating new knowledge through the processes of cognition, feeling and interaction.
3. The processes of data selection, data treatment, data pre-processing, and data mining are guided by the mental and social spaces in which they are performed. It is the previous knowledge and social constructions that determine what is (i) relevant data, (ii) the correct treatment of corrupted data, (iii) strategies for dealing with data fields, (iv) the number of rows and columns to reduce, and (v) the algorithm used in data mining.
4. The statistically relevant patterns and models resulting from data mining become meaningful to those interpreting them when they fit with previous knowledge. This happens even when people are faced by surprising or unexpected patterns. People perceive these patterns as surprising and valuable only in the face of the rich knowledge of what can be expected.
5. To be useful, the results of knowledge discovery should support the social construction of work relationships (for example, providing support to the co-ordination of tasks), shared physical artefacts (for example, models of data behaviour produced by data mining), shared social goals and projects (for

example, marketing campaigns guided by the results of data mining), and shared cultural norms and traditions (for example, integrating the culture of decision-making supported by knowledge discovery in databases). All these social constructions help make ideas and meanings tangible, support the negotiation of interests and facilitate communication between organisational participants. Thus, the whole organisation is involved in developmental cycles that lead to knowledge creation both at the individual and collective level.

6. Following on from 5 above, the interpretation of data mining results does not lead immediately to knowledge. The developed understanding needs to be tried out in organisational situations such as decision-making processes, organisational projects and the reformulation of practices. In this way, organisational participants create new ideas and mental models, reshape previous ones, redefine conceptual linkages and develop emotional responses towards ideas and models i.e. organisational participants create new experiences and relationships with their surrounding reality.

3 Test Case – Analysis of the Portuguese Army Database

This paper presents a test case in which a component of the Portuguese Army Database was analysed. This database, maintained by the Staff Command, records the personal data of the staff of the Army and the personal data of the young men (about 18 to 20 years old) that are examined for induction into military service. In Portugal military service is mandatory and each individual is given several tests, including medical, physical and intellectual tests. According to the results of the performed tests, each individual can be required, or not, to undertake military service.

Application domain understanding

Before the implementation of the several steps that integrate the KDD process, the Army Database was analysed in order to identify the relevance of each item (table) to the data mining objectives (tasks). Figure 2 shows that *Individual* represents the main item, which has relationships with the items *Parish, Municipality, District, Profession, Skill, Qualification, Idiom* and *Incapacity. Parish, Municipality* and *District* allow for the geographical characteristics of each *Individual*, while *Profession* indicates his or her main activity; *Skill* represents technical skills; *Qualification* the academic degree or years of study; *Idiom* the language(s) spoken and *Incapacity* the result of the various medical tests carried out.

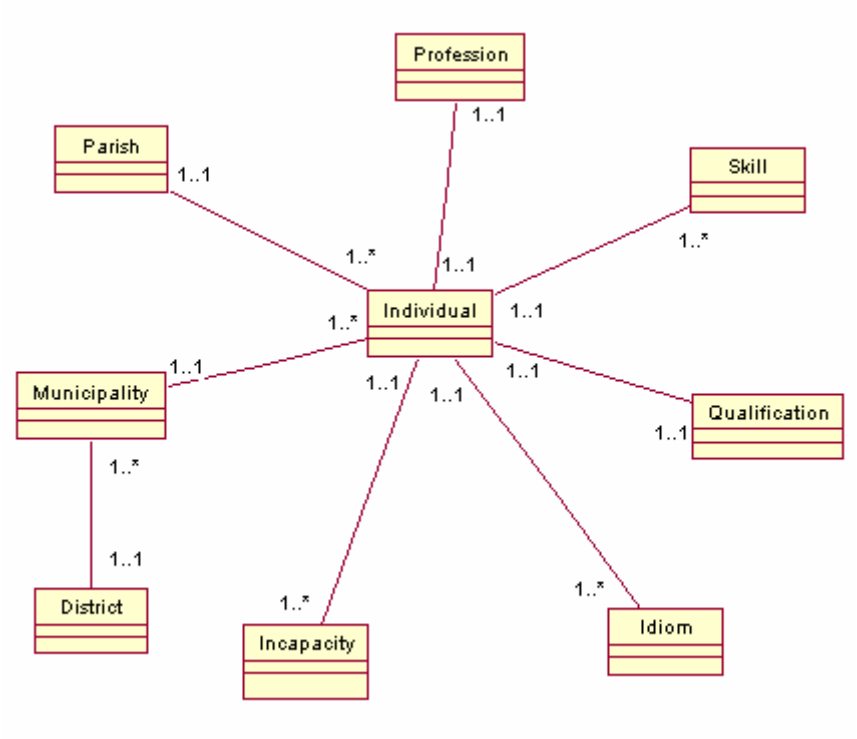

Figure 2 – A component of the Portuguese Army Database

For each table presented a set of data mining tasks was defined. Each one had as its purpose the implementation of the tasks associated with the process of knowledge discovery in geo-referenced databases *(Santos, 2001) (Koperski, Adhikary et al, 1996) (Han and Kamber, 2001)*. As in the Portuguese Army Database, most organisational databases integrate one dimension of data, the *geographical* (associated with addresses or postcodes), whose semantics are not used by traditional KDD systems. In order to include the geographical component in the analysis of the Portuguese Army Database, the principles of the PADRÃO system, which incorporates another step in the KDD process, were considered. This strategy and its principles are detailed and explained in Santos and Amaral *(Santos and Amaral, 2000)*. The PADRÃO system integrates a set of databases and principles of qualitative spatial reasoning that were implemented in the Clementine Data Mining system.

The task presented in this paper is concerned with the identification of spatial trends in the *Incapacity* attribute. A spatial trend *(Ester, Frommelt et al, 1998)* describes a regular change of one or more non-spatial attributes when moving away from a particular spatial object. In this task, one district of Portugal, the Braga district, will be considered and it will be determined if there exists an increasing or decreasing of the incapacity level when moving away from a given Municipality.

For the implementation of the defined data mining task, Detection of Spatial Trends, the Individual and *Incapacity* tables are required in addition to the table with the geographic component (*Municipality*). Below is presented the content of each of these tables, namely its attributes and its respective meaning, as well as the number of records stored in each table.

Individual Table: 1.328.573 records

Número	Long Integer	Military Identification Number
DataNascimento	Date	Date of Birth
Sexo	Char(1)	Sex
Concelho	Char(3)	Municipality Identification Code
Freguesia	Char(2)	Parish Identification Code
EstadoCivil	Char(1)	Marital Status

Municipality Table: 313 records

Concelho	Char(3)	Municipality Identification Code
DesigConcelho	Char(60)	Municipality Identification Name
DistritoAdm	Integer	District Identification Code

Incapacity Table: 492.204 records

Número	Long Integer	Military Identification Number
Lesão	Char(3)	Incapacity Identification Code
SIVAGE	Char(1)	SIVAGE Factor
GrauLesão	Integer	Level of incapacity

Analysing in detail the content of the *Individual* Table it is possible to verify that this database stores individuals born between 1858 and 1983. Up until 1981, inclusively, only Army staff was recorded. Before 1981 the personal details of the young men examined by the Army staff were also recorded in the database. This table contains 215 records with no date of birth. Four records present incorrect dates of birth, namely 12-06-979, 08-12-2886, 03-01-1989 and 14-03-1994. The Army Staff Command analysed these mistakes and indicated that the correct values should be 12-06-1979, 08-12-1886, 03-01-1889 and 14-03-1884, respectively. The unknown values were marked, in the data treatment phase of the KDD process, with the 'Nil' label.

Figure 3 shows the distribution of the individuals by *Municipality*, *Marital Status* and *Sex* (Concelho, Estado Civil and Sexo, respectively). Concerning *Marital Status*, the majority of the individuals have an *unknown* status (code '0', 60.56%), followed by *single* status (code 'S', 38.33%). The other records are distributed across the classes: *married* (code 'C'), *divorced* (code 'D'), *widower* (code 'V'), *judicial separated* (code 'J') and *marital union* (code 'M'). With respect to the *Sex* attribute, 1.319.175 instances are related to males and only 9.398 to females.

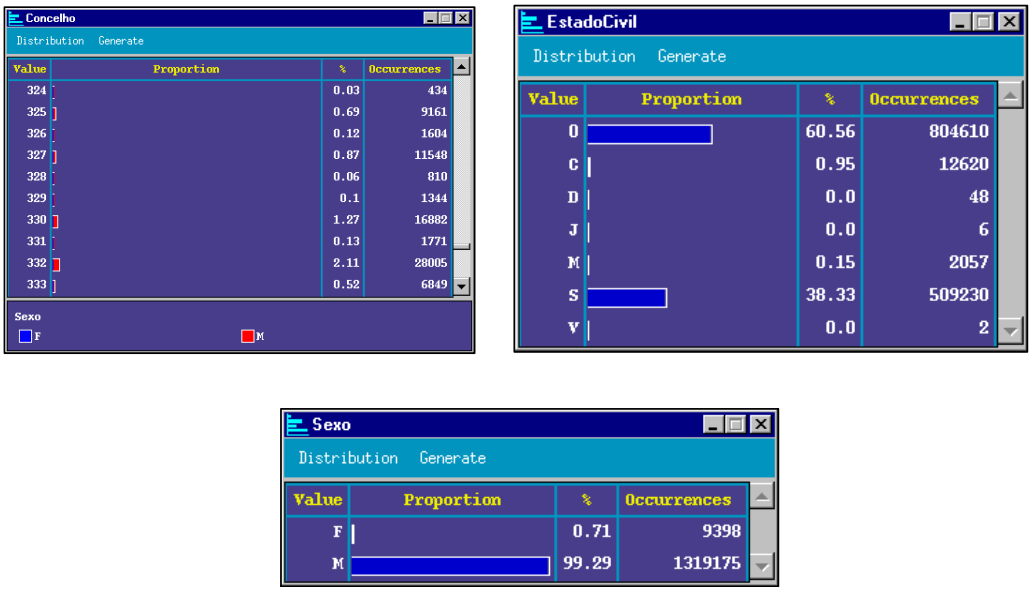

Figure 3 – Individual distribution by Municipality, Civil State and Sex

The *Incapacity* table stores the results of the medical tests carried out on the individuals. These results are coded by the SIVAGE factor, which includes six components that are: **S** (Arms), **I** (Legs), **V** (Vision), **A** (Audition), **G** (General State) and **E** (Emotional State). To each of these components the medical staff assigns an incapacity level according to the values: **1** (very bad), **2** (bad), **3** (normal), **4** (good) and **5** (very good).

Figure 4 presents a sample of the records stored in the Incapacity table, the distribution of the data according to the SIVAGE attribute and the histogram with the division of the various records by the different levels of incapacity. In the figure it can be seen that some records store values of the SIVAGE attribute that are not allowed (the correct ones are S, I, V, A, G and E). Since it is impossible to identify the real data values, the corresponding records will be removed from the database later.

In the *Incapacity* Table there are also some cases of duplication of information, which is verifiable in the recording of the components of the SIVAGE factor for a particular individual, as can be see in Figure 4 for the individual with the Identification Number (Numero) 73. These duplications originated from the examination of several medical tests that, when translated to a result, indicate that the same SIVAGE component sometimes has different levels. In these cases, and as suggested by the Army Staff Command, the record with the greatest degree of incapacity is taken, thus allowing an excess approximation of the reality.

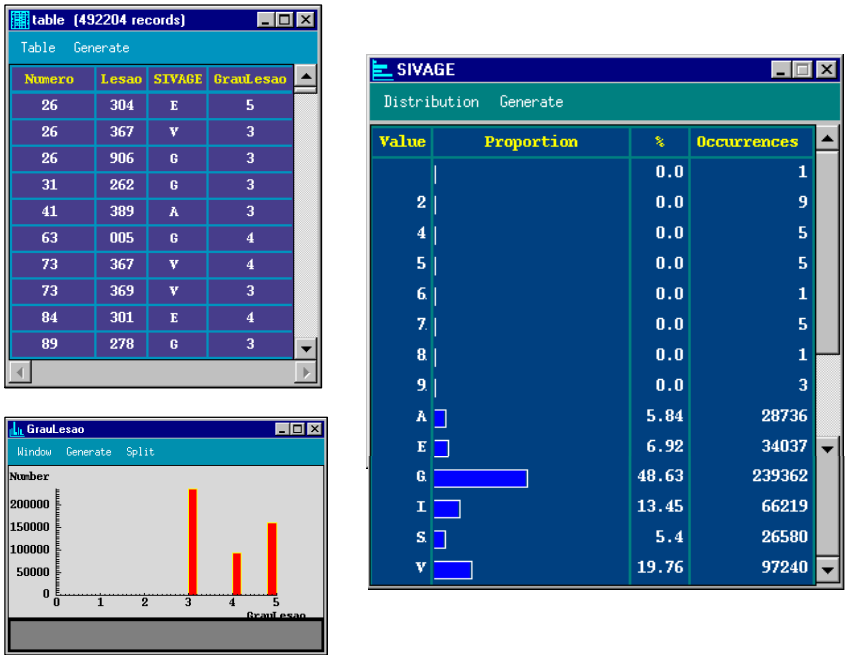

Figure 4 – Individual distribution by SIVAGE factor and Level of Incapacity

The KDD process

After the analysis of the available data it is possible to undertake the several steps of the KDD process. To the defined Data Mining task, detection of spatial trends in the Incapacity attribute, the relevant attributes and the attributes with incorrect values were identified.

The knowledge discovery application adopted was Clementine v5.2 of SPSS Inc. *(SPSS 1999)*. Clementine is a tool based on visual programming, which involves placing and manipulating icons representing processing nodes. The knowledge discovery process is defined in Clementine through the construction of a *stream* in which each operation on data is represented by a *node*.

For the defined task, the data selection, data treatment and data pre-processing steps were implemented in a single stream, minimising the time required for processing. With respect to the *Individual* Table:

- in the selection phase, the *Parish*, *Sex* and *Marital Status* attributes were excluded;
- in the data treatment phase:
  1. the dates of birth 08-12-2886, 14-03-1994, 12-06-979 and 03-01-1989 were rectified to 08-12-1886, 14-03-1894, 12-06-1979 and 03-01-1889, respectively;
  2. all records with unknown data of birth were labelled with '?'.

In the *Incapacity* Table, and concerning with the data selection phase, the Incapacity Identification Code was excluded, since it is implicit in the SIVAGE attribute. In the data treatment phase:

- all records with a SIVAGE factor different from S, I, V, A, G or E were deleted;
- for individuals with a duplicate SIVAGE factor the smallest one was deleted, as suggested by the Army Staff Command.

For the implementation of the defined task it was necessary to integrate the *Individual* and *Incapacity* Tables. In the resulting data the records geographically associated with the Braga district were selected. The 36.761 resulting records were randomly divided into two sets, the *training* and *test* sets. Figure 5 shows the stream that allowed the execution of the data selection, data treatment and data pre-processing phases and the construction of the *training set* (Tar4Treino.cdf) and the *test set* (Tar4Teste.cdf).

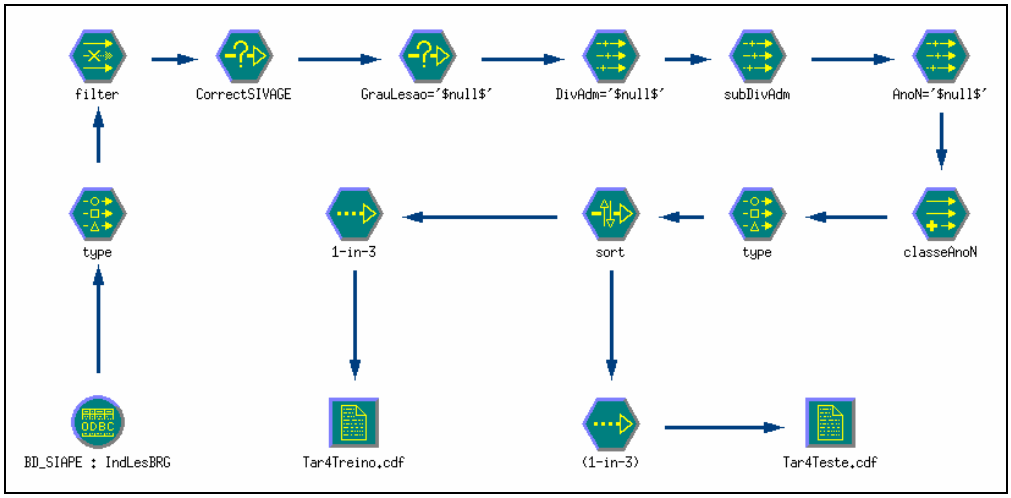

Figure 5 – Detection of spatial trends – *training* and *test* sets

As previously mentioned, the test case presented in this paper considered the knowledge discovery strategy of the PADRÃO system (Santos and Amaral, 2000). It allows the inclusion of the spatial component associated with the geo-referenced data analysed in the knowledge discovery process. PADRÃO allows for the inference of unknown spatial relationships and their use in the definition of a geographical model that can be included in the knowledge discovery process. Through this strategy it is possible to analyse geographical and non-geographical data simultaneously.

The knowledge available in the Spatial Knowledge Base of the PADRÃO system, which stores the reasoning rules that allow the inference of direction, distance and topological spatial relations, was used in the construction of three models, infDir, infDis and InfTop, which compliment each one in the spatial reasoning processes. With these models and as shown in Figure 6, it is possible to infer the unknown spatial relationships existing in the Municipalities of the Braga district. After the inferential process, the knowledge obtained is recorded in the Geographical Database (BDG:geoBraga) of the PADRÃO system.

In the Data Mining phase it is only necessary to select the appropriate algorithm to carry out the defined task. The detection of spatial trends has the identification of regular changes in the incapacity level of several individuals as its objective. In order to do so, it is necessary to integrate the geographical component of the Braga district with the *training set* (previously obtained). This process is highlighted in Figure 7. The algorithm selected was C5.0 that allows the introduction of decision trees. The generated rules describe spatial tendencies in the analysed data. For code V (*vision*) of the SIVAGE factor there exists a distribution of the different incapacity levels in the various Municipalities of Braga.

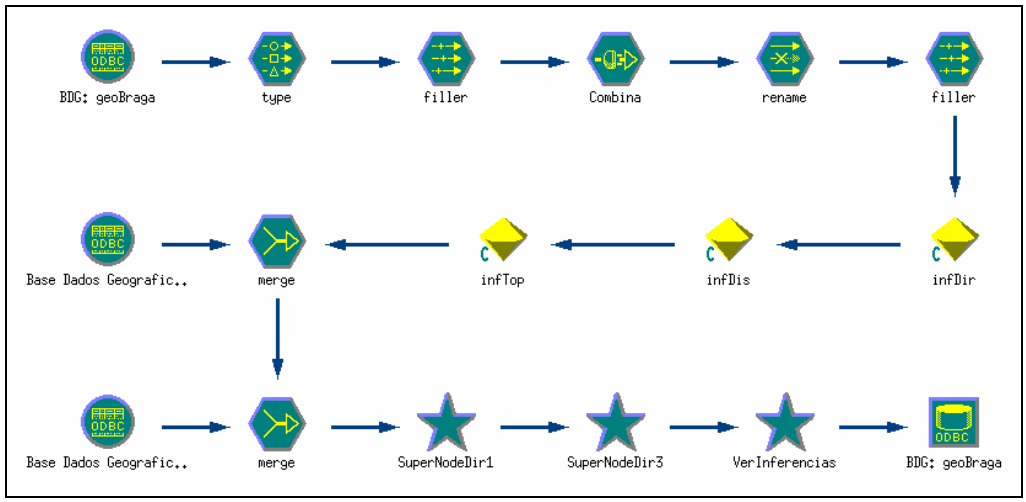

Figure 6 – Inference of unknown spatial relationships

Selecting a particular *Municipality*, Guimarães with the code 308, and analysing the qualitative distances[1] existing between the other Municipalities of the district and the *Municipality* with the code 308, it was possible to identify a small set of rules that explicitly express a diminution of the incapacity level when moving away from the *Guimarães* Municipality. The model obtained (GrauLesaoGUIM) is shown in the left hand lower corner of Figure 7. For *Municipalities* closest (code p) to Guimarães the incapacity level is 5, while for *Municipalities* far (code d) from Guimarães this level is 3.

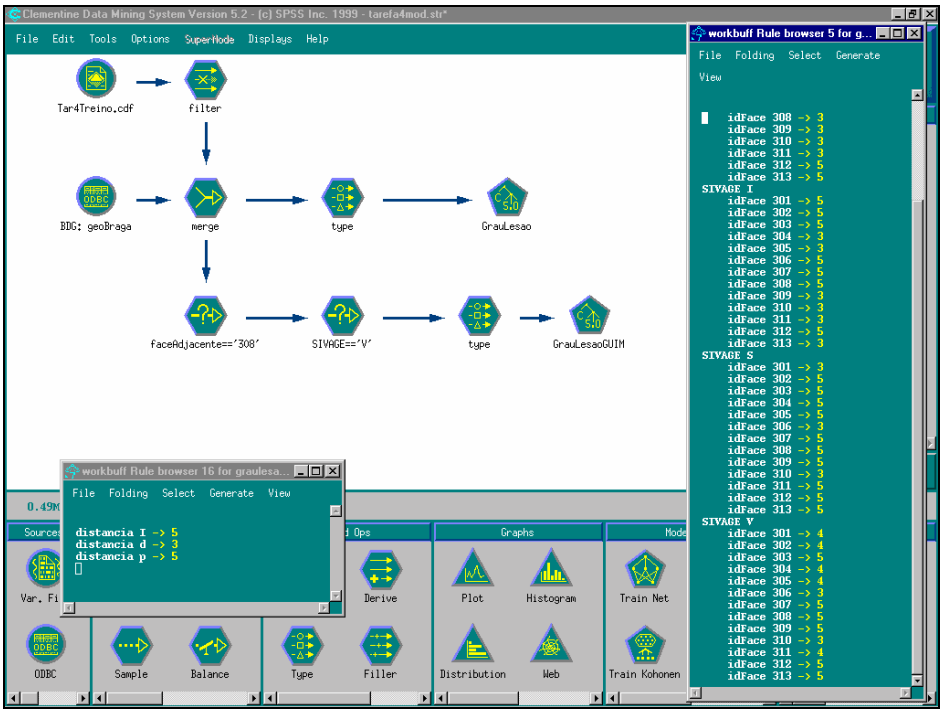

Figure 7 - Spatial trends in the SIVAGE factor

[1] The qualitative distances considered were very close (mp), close (p), far (d) and very far (md).

The *test set* is used in the interpretation of results phase to verify the confidence in the models built in the Data Mining phase. With respect to the model previously obtained, Figure 8 shows a percentage of confidence of 47.17%. Although it is a low value it is thought that it is caused by the poor distribution of data in the several classes available. At this stage it is possible to return to a previous phase, modify some decisions and proceed with the knowledge discovery process. This hypothesis was not explored, since it constitutes a repetition of the process presented up until now.

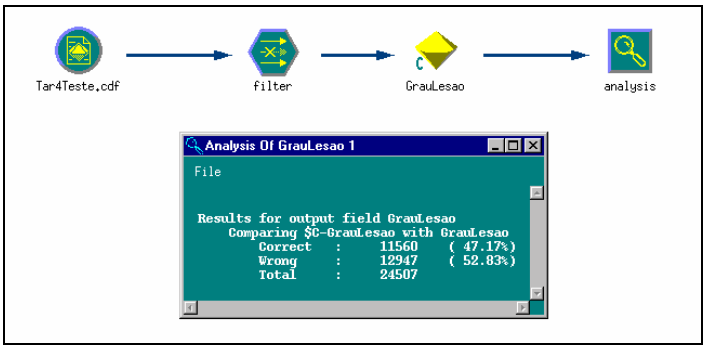

Figure 8 – Confidence of the identified model

The PADRÃO system permits the visualisation of the results of the knowledge discovery process as a map. In this system the several rules that integrates a model are recorded in the Patterns Database *(Santos and Amaral, 2000)*. At the same time the user has the possibility of running the VisualPadrão tool (an application implemented in Visual Basic that was integrated into the environment of the Clementine toolkit) and visualising the desired model as a map (Figure 9).

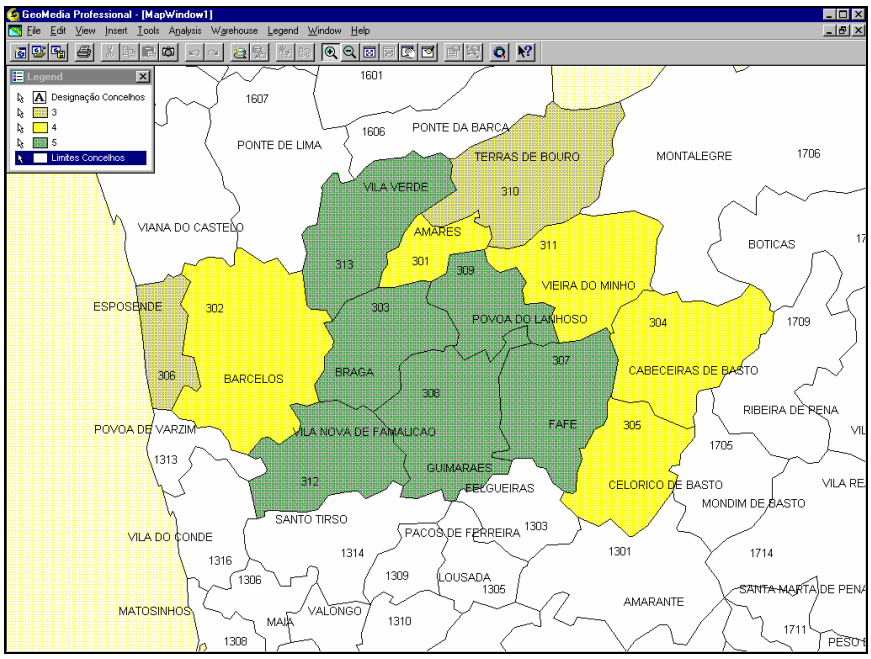

Figure 9 – Spatial tendencies in code V of the SIVAGE factor

4 Discussion

Since section 3 describes a test case and not a use of the PADRÃO System in a real organisational setting, the discussion presented here is limited to general issues related with Data Mining of the Portuguese Army Database in order to identify spatial trends in the *Incapacity* attribute.

The initial choice of the attribute to use for identifying spatial trends is guided by the interests of the stakeholders of the process, which are rooted in the previous work experience. The Army Staff Command had an interest in identifying the spatial tendencies of the visual incapacities. It is known that visual incapacities are often the reason that impedes call up for military service, which is often the cause of problems for young men who have started their first job. Thus, there have been many attempts to fake a lower degree of visual incapacity in order to avoid being called up for military service. Thus it was thought that the spatial tendencies in code V of the SIVAGE factor would help to detect potential frauds by leading to more detailed medical tests.

It is worth noting that other codes hold very subjective information. This is the case for code G (general state) and especially code E (emotional state). An analysis of the spatial data model produced with this subjective information may produce very different knowledge among staff (i) with a long experience in the selection of young men and the culture of the Army who support the classification of the medical tests into the five levels incorporated into the Level of Incapacity field of the Incapacity Table, or (ii) inexperienced in the task or culture that render the codes meaningful.

To make the results produced by the PADRÃO system useful, it is necessary to address several aspects, namely:

1. It is important to be aware of the potential pre-conceptions that are introduced in the process of subjecting young men to medical tests. In the specific case presented in this paper, should the Army Staff Command suspect fraud when the medical tests show a lower value in the Level of Incapacity field than that usually found in young men from the same municipality? Should such young men be treated with suspicion?
2. Formal and/or informal orientations need to be established in order to define how the spatial models will influence decisions about inducting young men subjected to the medical tests.
3. An evaluation must be performed in order to define what measures must be taken to prevent unfair treatment of those young men that are inducted and show an acceptable low value in the level of incapacity field.

Also, it is important to define the structures of the Portuguese Army that can benefit from information provided by spatial trends that can be found in the Army Database and to find effective formal and informal working relationships that enable the sharing of results of Data Mining, experimentation with those results and sharing of the insights gained.

5 Conclusions and Future Work

This paper seeks to integrate the process of knowledge discovery in databases into the wider context of the creation and sharing of organisational knowledge. To achieve this goal the view of the researchers of the knowledge discovery process and its integration in the social process of knowledge creation in organisations is first presented. Then, came a test case carried out to test the PADRÃO system, which integrates a set of databases and principles of qualitative spatial reasoning that were implemented in the Clementine Data Mining system. Finally, the actual process, its results and an interpretation of them according to the core ideas of the social construction of knowledge is presented.

The grounding of the discussion in a test case and not a case study, limited the ability to fully highlight the richness of the insights resulting from the integration of the KDD concepts and practices into a widely studied sociological perspective of the social construction of organisational reality *(Kafai, 1996)*. In the near future, the authors expect to be able to provide these insights by carrying out case studies and action research. The authors have planned a project to study the contribution of KDD to the creation of shared knowledge and the reformulation of work practices in a Portuguese hospital. The aim is to define guidelines for designing and implementing KDD processes that make a clear contribution to organisational efficiency.

6 Acknowledgements

We thank the Portuguese Army for making the Staff Command's database available for analysis. We thank Anthony Lavender for his help in improving the English writing of this paper.

7 References

Ester, M., A. Frommelt, et al. (1998). Algorithms for Characterization and Trend Detection in Spatial Databases. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, AAAI Press.

Fayyad, U. and R. Uthurusamy (1996). "Data Mining and Knowledge Discovery in Databases." Communications of the ACM 39(11): 24-26.

Han, J. and M. Kamber (2001). Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.

Kafai, Y. and M. Resnick, Eds. (1996). Constructionism in Practice: designing, thinking, and learning in a digital world. Mahwah, New Jersey, Lawrence Erlbaum Associates.

Koperski, K., J. Adhikary, et al. (1996). Spatial Data Mining: Progress and Challenges. Proc. of the 1996 SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, Montreal.

Ramos, I., João A. Carvalho (Accepted for publication). "Towards constructionist Organizational Data Mining (ODM): changing the focus from technology to social construction of knowledge", In Barko and Nemati (Eds.), Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance, Hershey, PA: Idea Group (to be published soon).

Santos, M. and L. Amaral (2000). Knowledge Discovery in Spatial Databases through Qualitative Spatial Reasoning. PADD'00 Proceedings of the 4th International Conference and Exhibition on Practical Applications of Knowledge Discovery and Data Mining, Manchester, 11-13 April.

Santos, M. and L. Amaral (2000). “Knowledge Discovery in Spatial Databases: the Padrão's qualitative approach.” Cities and Regions GIS special issue (November): 33-49.

Santos, M. Y. (2001). PADRÃO: Um Sistema de Descoberta de Conhecimento em Bases de Dados Geo-referenciadas, PhD Thesis (in Portuguese), University do Minho.

SPSS (1999). Clementine, User Guide, Version 5.2, SPSS Inc.